# Practical 1: Linear and logistic regression, random and fixed effect

Garyfallos Konstantinoudis

Spring Term 2025

## Part 1: Analysing type 2 diabetes progression using linear regression

Type 2 diabetes is a long-term metabolic disorder that is characterized by high blood sugar, insulin resistance, and relative lack of insulin. The number of people diagnosed with diabetes in the UK has more than doubled in the last twenty years. According to diabetes.org.uk/ figures show that there are now almost 3.7 million people living with a diagnosis of the condition in the UK, an increase of 1.9 million since 1998.

In this practical we consider an observational study that measures progression of type 2 diabetes (quantitative score) as outcome and several clinical parameters like age. sex, and bmi, but also common risk factors like map: blood pressure, tc: total cholesterol, ldl: low-density lipoprotein, hdl: high-density lipoprotein, tch: total cholesterol, ltg: triglycerides, and glu: glucose. The dataset includes n=442 cases. It is available in the lars package, which is easy to install using for example the install.packages("lars") command.

As a first step we load the data and assign x as a data.frame of the predictors and y as the quantitative score of type 2 diabetes progression.

```r
library(lars)
data(diabetes)
x = as.data.frame.matrix(diabetes$x)
y = diabetes$y
```

From the literature we know that the following 6 predictors are important for type 2 diabetes progression: sex, age, bmi, glu, map and ltg. We consider these 6 predictors as model 1.

Question 1.1 Look at the correlation structure between those 6 predictors and discuss the implications. Use the function corrplot() in the corrplot package to visualise the correlation structure.

Question 1.2 Does the outcome disease progression follow a Normal-distribution? Look at general summary statistics of y, plot a histogram and a q-q plot against the Normal-distribution.

Question 1.3 Fit a linear model including the predictors of model 1 ( sex, age, bmi, glu, map and ltg) using the lm() function and discuss the summary of the model.

Question 1.4 Perform model diagnostics and outlier detection of model 1. Do you think this is a good model fit? Justify your answers.

Question 1.5 Compute the OLS regression coefficient estimate using matrix multiplication $\hat{\beta}_{OLS} = (x^t x)^{-1} x^t y$. Use the solve() function to invert a matrix. R distinguishes between scalar multiplication ($*$) and matrix multiplication (%$*$%). Make sure to use the matrix multiplication (%$*$%) for this task and ensure that your matrices have the correct dimensions. Add an intercept by including a row of ones like for example this.

```r
x1 =  cbind(rep(1,nrow(diabetes$x)), x$sex, x$age, x$bmi, x$glu, x$map, x$ltg)
```

Question 1.6 Compute the regression coefficient estimate using the sample covariance based estimate, defined as $\hat{beta}_{COV} = cov(x)^{-1}cov(xy)$. Use the solve() function to invert the covariance matrix cov(x) of dimension 6 x 6 and compute this estimate without the intercept using

```
x11 =  cbind(x$sex, x$age, x$bmi, x$glu, x$map, x$ltg)
```

Question 1.7 Compare the 3 estimates from questions 1.3, 1.5 and 1.6.

Question 1.8 Fit model 2 that only includes glucose and compare how it differs from the multivariable model 1.

## Part 2: Predict type 2 diabetes progression using linear regression

Assume we only observed the first 300 cases and use these cases as training data.

```
x_train = data.frame(x[1:300,])
y_train = y[1:300]
```

Now we can consider the remaining 142 cases as new data points for whom we want to predict disease progression.

```
x_new = data.frame(x[301:442,])
y_new = y[301:442]
```

Question 2.1 Use the linear model 1 to predict the disease progression for the 142 cases with predictor information stored in x_new.

Question 2.2 Evaluate the error of your prediction based on linear model 1 by computing the squared difference between the predicted progression and the actual observed progression saved in y_new. Plot a histogram of the squared difference and compute the mean and median.

Question 2.3 Repeat the steps 2.1 and 2.2 using the univariable linear model 2 including only glucose. Contrast the prediction error of linear model 2 with the prediction error of linear model 1.

Question 2.4 Is it good practise to evaluate the prediction performance on a single training data? How appropriate is the split to take the first 300 cases?

## Part 3: Distinguishing between severe and mild cases of type 2 diabetes using logistic regression

Doctors are particularly concerned with type 2 diabetes cases that have a bad disease progression, in particular cases that have a disease progression score larger than 200. Binarise your outcome like this:

```
y_binary = as.numeric(y>200)
```

Question 3.1 Fit a generalised linear model using the glm() function that can distinguish between bad disease progression and normal progression. Use the 6 predictors as considered in model 1. Look at the summary of the glm output and interpret the findings.

Question 3.2 Consider now model 2 including only glucose. Fit a glm and see if glucose can distinguish between bad disease progression and normal progression.

Question 3.3 Look again at the training data (x_train and ybin_train) based on the first 300 cases, where

```r
ybin_train = y_binary[1:300]
```

Build a prediction rule based on model 1 using the training data (x_train and ybin_train) using the glm function. In a second step predict which of the new samples (using x_new as predictor matrix) are at high risk for having a bad diagnosis. How many of the 142 new observations have a probability larger than 0.5 to have bad progression?

PS Use the inverse logit function ($logit^{-1}(eta) = exp(eta)/(exp(eta) + 1)$) to transform the linear predictor ($eta = x\beta$) back to a probability which ranges between 0 and 1.

## Part 4 (Optional): Which risk factors are important for type 2 diabetes progression?

Look again at the complete dataset including all n=442 cases and all 10 predictors. How would you perform variable selection to decide which variables are important for disease progression in type 2 diabetes?

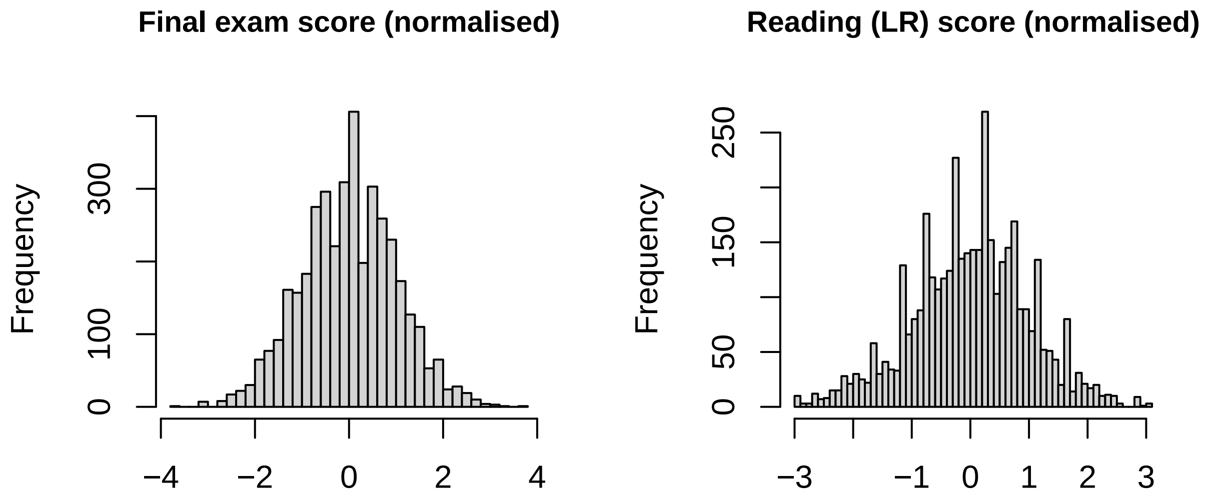
## Part 5: Linear mixed model: Exam scores from London

This section considers exam scores of 3,935 students from 65 schools in Inner London. In particular, we want to find out how the final exam score can be predicted by reading abilities as measured in the London reading (LR) test. Please adjust the path to the dataset according to your computational setup.

```r
load("../data/exam.London")
dim(exam)
```

```
## [1] 3935   10
```

```r
par(mfrow=c(1,2))
hist(exam$normexam, breaks =50, main="Final exam score (normalised)", cex.main=0.9, xlab="")
hist(exam$standLRT, breaks =50, main="Reading (LR) score (normalised)", cex.main =0.9, xlab="")
```



```r
dev.off()
```

```
## null device
##           1
```

Additional covariates of the data are:

- school: School ID - a factor
- schgend: School gender - a factor. Levels are 'mixed', 'boys', and 'girls'
- schavg: School average of intake score
- vr: Student level Verbal Reasoning (VR) score band at intake - 'bottom 25%', 'mid 50%', and 'top 25%'
- intake: Band of student's intake score - 'bottom 25%', 'mid 50%' and 'top 25%'
- sex: Sex of the student - levels are 'F' and 'M'
- type: School type - levels are 'Mxd' and 'Sngl'
- student: Student id (within school) - a factor

Question 5.1

Fit a linear model to test if there is a linear relationship between reading ability and the final exam score and plot a scatterplot of exam score against reading ability.

Question 5.2

Are there any potential issues with the standard linear model?

Question 5.3

Fit a fixed effect model accounting for the effect of schools using the lm() function where you add school (as.factor()) as covariate. What is the interpretation of the model and how many additional parameters do we need to estimate?

Question 5.4

Now use the function in the lme function in the

```r
library(nlme)
```

package to estimate a random effects model with a random intercept depending on the school. What is the interpretation of the fixed effect? How many parameters do we need to estimate compared to the fixed effects model?

Question 5.5

What is the intra-class correlation coefficient for this model (lecture 1c, slide 38-40) and how do you interpret it?

Question 5.6

Add a random slope depending on school to your model and see if the effect of the fixed effects changes.

Question 5.7

Which of the covariates are individual-level and which are group-level variables? Re-fit your random intercept model adding the group-level variables to the random effects model.

Question 5.8

Compare the random intercept (Q5.4) and the random intercept and slope model (Q5.6) using the likelihood ratio test and discuss which one has the better model fit.

Question 5.9

Compare the random intercept (Q5.4) and the one with the additional covariate (Q5.7) using the AIC and BIC (note that those two models are not nested) and discuss which one has the better model fit.

## Part 6: Linear mixed model: Survival on the Titanic

The sinking of the titanic was one of the greatest disaster in navel history. After colliding with an iceberg, the titanic sank and 1,502 out of 2,224 passengers and crew were killed. The following data set has collected information on n=1,309 of the passengers and their survival.

```
titanic = read.csv("../data/titanic.csv")
dim(titanic)
```

```
## [1] 1309   14
```

```
table(titanic$survived)
```

```
##
##   0   1
## 809 500
```

The dataset includes:

- survival: Survival (0 = No; 1 = Yes)
- class: Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- name: Name
- sex: Sex (1=female, 2=male)
- age: Age
- sibsp: Number of Siblings/Spouses Aboard
- parch: Number of Parents/Children Aboard
- ticket: Ticket Number
- fare: Passenger Fare
- cabin: Cabin
- embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
- boat: Lifeboat (if survived)
- body: Body number (if did not survive and body was recovered)

For more information on the data and a data challenge called 'Machine Learning from Disaster' see

https://www.kaggle.com/c/titanic

In the following we want to test if the phrase 'women and children first' was adapted for the evacuation of the titanic.

Question 6.1

Since survival is a binary outcome here, use a glm to test if age and sex had an effect on survival.

Question 6.2

Next step is to account for the passenger class (variable pclass) in a fixed effects model and discuss the implications and difference to the simple model.

Question 6.3

Discuss whether to include the passenger class as a fixed or random effect and fit a random effects model with a random intercept depending on passenger class using the glmer() function in the

```r
library(lme4)
```

package.

Question 6.4

Add a random slope depending on passenger class to your model and compare it with the random intercept only model using a likelihood test.

Question 6.5

How do you explain the difference in results after accounting for passenger class? Use a boxplot and a violin plot for age depending on passenger class to illustrate your argument.